Sekwencjonowanie genomu ma bezprecedensowy wpływ na nasze zrozumienie gatunków żyjących na Ziemi. Koszt sekwencjonowania pierwszego prawie kompletnego genomu ludzkiego (ukończonego w 2001 r.) wyniósł około 3 miliardy dolarów. Po 20 latach zmalał on do jedynie około 1 tys. USD. To ogromna poprawa i możemy się spodziewać, że to nie koniec redukcji kosztów.

Kluczowymi graczami na rynku są Illumina, Oxford Nanopore i PacBio. Ich sekwenatory znacznie się od siebie różnią. Illumina oferuje sekwencjonowanie o wysokiej przepustowości i generuje krótkie odczyty wysokiej jakości (100–250 par zasad). PacBio generuje znacznie dłuższe (np. 15 kbp) odczyty wysokiej jakości, ale przepustowość jest niższa. Wreszcie Oxford Nanopore daje jeszcze dłuższe odczyty (np. 1Mbp), ale jakość jest umiarkowana. Biorąc to, co najlepsze z nich wszystkich, możemy asemblować de novo nawet złożone genomy, takie jak ludzkie.

Istnienie takich stosunkowo tanich technologii zaowocowało uruchomieniem kilku projektów sekwencjonowania całych genomów na dużą skalę. Niedawno konsorcjum Telomere-to-Telomere opublikowało pierwszy kompletny ludzki genom. Wypełnienie wszystkich luk w sekwencji ludzkiego genomu od 2001 roku, kiedy opublikowano pierwszą sekwencję genomową ludzkiego organizmu, zajęło 20 lat, co pokazuje, jak bardzo złożone było to zadanie. Innym przykładem jest Human Pangenome Project (HPP). Jego celem jest zasemblowanie de novo setek ludzkich genomów, zapewniając dużą bazę danych różnorodności genomowej w ludzkich populacjach. Do czerwca 2022 roku opublikowano prawie sto wysokiej jakości sekwencji ludzkich genomów. Takie wysiłki powinny zmniejszyć niedostateczną reprezentację wielu sekwencji genomowych w obecnie używanym genomie referencyjnym. Oczekuje się, że wykorzystanie wielu genomów z różnych populacji znacznie poprawi określanie wariantów genomowych w medycynie spersonalizowanej.

Realizowane są również projekty skupiające się na zbieraniu kompletnych genomów różnych gatunków. Jednym z przykładów jest Vertebrate Genomes Project. Jego celem jest dostarczenie wysokiej jakości genomów referencyjnych dla wszystkich ok. 70 000 zachowanych gatunków kręgowców. Cele Earth BioGenome Project (EBGP) są również bardzo ambitne. Planuje się zsekwencjonowanie wszystkich znanych gatunków eukariotycznych w ciągu 10 lat.

Jednym z głównych celów sekwencjonowania genomu jest identyfikacja genów kodujących białka. Pierwsze duże bazy danych białek opublikowano w latach 90. XX wieku. Obecnie prawdopodobnie najpopularniejsza z nich, Pfam, zawiera łącznie prawie 20 tys. rodzin białek i łącznie 169 milionów sekwencji białek. Największa rodzina zawiera prawie 3 miliony sekwencji.

W niniejszym projekcie skupimy się na opracowaniu nowych algorytmów dla wspomnianych danych. Po pierwsze, aby uwzględnić koszty przechowywania i przesyłania ogromnych zbiorów danych, stworzymy specjalizowane algorytmy kompresji. Po drugie, opracujemy algorytmy wykorzystania zebranego genomu i sekwencji białkowych do rozwiązania wybranych problemów biologicznych.